Empirical methods for the estimation of the mixing probabilities for socially structured populations from a single survey sample

by

Stephen P. Blythe
Carlos Castillo-Chavez
George Casella

BU-1072-MA

April 1990
Revised October 1991

Abstract

The role of variability of sexual behavior in the transmission dynamics of HIV and AIDS has been illustrated, through the use of mathematical models, by several investigators. Models that capture heterogeneities due to rates of sexual partner change, changing behaviors, and demographic factors are invaluable in the study of the dynamics of sexually-transmitted diseases. Models that describe the processes of mixing between individuals and/or pair formation/dissolution have been formulated in great generality by Blythe, Busenberg, and Castillo-Chavez. Mathematical formulas describing all forms of mixing for one- and two-sex populations as structured deviations from random mixing have been obtained by Busenberg and Castillo-Chavez. In this paper we describe some practical methods for estimating the deviations from random mixing from a single survey sample. This method can be potentially very useful if one considers the difficulties – technical and political – involved in the gathering of behavioral and mixing data. We include a description of the role of the estimated mixing probabilities in models for the spread of HIV, a discussion of alternatives and possible extensions of the methods described in this article, and an outline of future directions of research. We note that despite the fact that the mixing probabilities $\{p_{ij}(t)\}$ are time-dependent, we are able to make use of time-independent parameters – the matrix of constant quantities $\{\phi_{ij}\}$ which are related to the initial deviations from random mixing – in the estimation of the dynamic mixing probabilities $\{p_{ij}(t)\}$.

Key words: HIV, AIDS, proportionate mixing, non-random mixing, pair-formation, estimation, sexually-transmitted diseases

1. INTRODUCTION

Projections and prediction of future trends of HIV and AIDS incidences cannot be made with confidence due to the many uncertainties involved in the measurement of key epidemiological and sociological parameters (Anderson 1988; Gupta *et al.* 1989; Anderson *et al.* 1989, 1990; Schwager *et al.* 1989; Castillo-Chavez *et al.* 1989; Blythe and Castillo-Chavez 1989). Methods involving some form of extrapolation and back-calculation have provided useful ways of obtaining short-term projections for the future temporal trends in AIDS and HIV incidence (Brookmeyer and Gail 1988; Cox and Medley 1989; Isham 1989; Day *et al.* 1989; Karon *et al.* 1988, 1989). Current projections of HIV prevalence, for example, depend largely on two factors: the shape of the incubation period distributions and the temporal patterns of HIV incidence of distinct interacting subpopulations. Although the uncertainties involved in estimating incubation period distributions for different groups are being reduced, albeit slowly, many problems associated with the estimation of incidences (new cases of infection per unit time) are still unresolved. The lack of sufficiently complete longitudinal serological and behavioral data suggests that our understanding of the consequences of these factors in disease dynamics may have to rely on experimental "data" generated by transmission dynamics models that incorporate realistic and potentially measurable social structures (Castillo-Chavez *et al.* 1989b,c; Blythe and Castillo-Chavez 1989; Castillo-Chavez and Blythe 1989; Sattenspiel and Castillo-Chavez 1990; Huang 1989; Huang *et al.* 1992), and in the development of methods that make full use of cross-sectional data. This paper *begins* to address the latter issue.

Model methodology has improved considerably over the last few years (see Castillo-Chavez 1989 for a review of this literature), yet much work remains to be done (Sattenspiel and Castillo-Chavez 1990; Castillo-Chavez 1989; Busenberg and Castillo-Chavez 1989, 1991; Castillo-Chavez *et al.* 1991). Models incorporating age-structure, variable infectivity, long and variable periods of infectiousness, risk levels, vertical transmission, and other factors have been developed (Anderson 1988; Schwager *et al.*

1989; Sattenspiel and Castillo-Chavez 1990; Castillo-Chavez 1989; Castillo-Chavez *et al.* 1989*b*) but important questions raised by the inherent limitations of some modeling approaches still remain. To reduce the effects of these limitations, we need to determine ways of comparing results *across* models (Sattenspiel and Castillo-Chavez 1990; Castillo-Chavez *et al.* 1991). For example, numerical studies on a variety of models suggest that heterogeneity in sexual behavior is a very important factor in the transmission dynamics of HIV; however, we need to have a ranking of the effects of these heterogeneities in HIV transmission. To answer questions of this type, we need more data and a better understanding of the principles and assumptions underlying different modeling approaches. The lack of political support for large scale surveys of sexual behavior in the general population means that in the foreseeable future, we will rely mostly on mathematical models for qualitative and quantitative evaluation of the effects of heterogeneity in its different forms.

There are numerous alternative ways of describing heterogeneity (Blythe and Castillo-Chavez 1989, 1990;Anderson *et al.* 1990;, Karon *et al.* 1989; Sattenspiel and Castillo-Chavez 1990; Busenberg and Castillo-Chavez 1989, 1991; Castillo-Chavez *et al.* 1991; Blythe and Anderson 1988; Castillo-Chavez and Busenberg 1991), but in this paper we will be concerned exclusively with the development of practical methods of estimating parameters that aim at the heart of the question of "who mixes with whom," from a single (cross-sectional) survey sample. To this end we have divided this article into four parts. In the Section 2, we describe a multigroup one-sex model that arises in the study of the transmission dynamics of HIV. The fact that this model incorporates very general forms of mixing allows us to explain in general terms, in the Section 3, the general estimation problem associated with the mixing parameters. In Section 4, we describe our *empirical* approach to the estimation of the mixing parameters, and in Section 5 we provide some numerical examples and explain the methodology used. The paper concludes with a discussion of future applications of this approach, alternative approaches, and research directions. Relevant technical information is collected in two appendices.

2. BASIC TRANSMISSION MODEL FOR HIV DYNAMICS

In order to discuss the problem of estimating the parameters associated with the mixing/pair-formation process we introduce a model for the spread of HIV/AIDS that focuses on these processes. The detailed model is provided in Appendix A; here we concentrate on describing a key component of this type of model, namely the incidence rate (new cases of infection per unit time). The mixing probabilities, as well as other behavioral and epidemiological parameters, determine the rate at which new infections are generated. The incidence rate is given by a nonlinear function of the different interacting subpopulations, and it is in the context of this expression that we will describe our empirical estimation procedure.

To illustrate the procedure, we consider a population of homosexually-active individuals (the two-sex case can also be addressed). The population is divided into classes or subpopulations, where such classes can be identified by race, socio-economic background, average degree of sexual activity, etc. For more general models that take into consideration factors such as chronological age, age of infection, variable infectivity, sex, and partnership duration the reader is referred to the work of Busenberg and Castillo-Chavez (1989, 1991) and Castillo-Chavez et al. (1991); for the most up-to-date mathematical analysis of this type of models see (Castillo-Chavez *et al.* 1989b,c; Huang, 1989; Cooke *et al.* 1991; Huang *et al.* 1992). The N sexually active subpopulations are divided into three epidemiological classes: S_i (susceptible individuals), I_i (HIV-seropositive asymptomatic or with mild symptoms), and A_i (HIV seropositive, with severe symptoms) for i=1,$\cdots$,N. We assume that only S- and I-individuals are sexually active, and the sexually-active populations are denoted by $T_i(t) = S_i(t) + I_i(t)$, i = 1,...,N. $B_i(t)$ denotes the i^{th} incidence rate at time t, that is, the number of new infective cases in subpopulation i per unit time. $B_i(t)$ is a complicated function that depends on the frequency and type of sexual interactions that susceptible individuals in group i have with all other individuals (including those in group i).

To describe the expression for the i^{th} incidence rate we need more definitions: β_j denotes the transmission rate per infective group j partner (alternative definitions for this parameter are available, see Castillo-Chavez *et al.* 1989b; Cooke *et al.* 1991), C_i denotes the average number of new partnerships per unit time of group i individuals, and $p_{ij}(t)$ denotes the fraction of partnerships of individuals in group i with individuals in group j or, equivalently, the probability that a group i individual will mix with a group j individual. Since $C_iS_i(t)$ denotes the "average" number of partnerships per unit time formed by susceptible individuals in group i, $C_iS_i(t)p_{ij}(t)$ denotes the *average mixing rate* group i susceptibles with group j individuals, and $C_iS_i(t)p_{ij}(t)I_j(t)/T_j(t)$ denotes the *average mixing rate* with group j infectives. Multiplying this last expression by β_j we obtain the average rate at which partnerships with j infectives lead to new i infectives. Summing over all groups (j = 1,..., N) we obtain the total average rate of infection in group i, that is, the number of new cases of infection per unit time in group i generated by the interactions of group i susceptibles with infectives of all other groups. This time-dependent rate is prescribed by the mixing matrix $\{p_{ij}(t)\}$ (see Table 1). A summary of notation, and the explicit expression for the i^{th} incidence rate $B_i(t)$, appears in Table 1. The full dynamic model is described by specifying the rates of change, per unit time, of all the epidemiological classes. The formulae are provided in Appendix A.

The main objective of this paper is to specify ways of estimating the mixing probabilities $p_{ij}(t)$. Because the transmission-dynamic model given by equations (A1)-(A3) is deterministic, the specification of the initial state of the system (i.e. the number of susceptible, infectives, and the incidence at time t=0) uniquely characterizes all future states (i.e., all future population sizes of the epidemiological classes $S_i(t)$, $I_i(t)$, $A_i(t)$, as well as the sizes of the incidences $B_i(t)$). Specifically, knowledge of $S_i(0)$, $I_i(0)$, and the N^2 quantities $p_{ij}(0)$ uniquely determines the course of the model epidemic provided that we have specific formulas for the $p_{ij}(t)$'s. The mixing probabilities $p_{ij}(t)$ (i,j = 1,2,3,..N) depend on

several factors, and generally are given by complicated functions of the sizes $T_i(t)$ of the N subpopulations and the necessary behavioral and epidemiological parameters. Hence to forecast the state of the model epidemic at all future times, we need to have an explicit functional form for these mixing probabilities. Such functional forms have generally been selected in some ad hoc manner; in the next section we describe a systematic approach to mixing probability estimation.

3. ESTIMATION: FORMULATION FOR THE MIXING/PREFERENCE MATRIX

The mixing inter- and intra-group probabilities $p_{ij}(t)$'s must satisfy the following properties at *all* times:

$$0 \leq p_{ij}(t) \leq 1, \qquad i, j = 1,\cdots,N, \tag{i}$$

$$\sum_{j=1}^{N} p_{ij}(t) = 1, \qquad i = 1,\cdots,N, \tag{ii}$$

$$C_i T_i(t) p_{ij}(t) = C_j T_j(t) p_{ji}(t), \qquad i, j = 1,\cdots,N. \tag{iii}$$

$$C_i T_i(t)\ C_j T_j(t) = 0 \ \Rightarrow \ p_{ij}(t) = p_{ji}(t) = 0. \tag{iv}$$

Properties (i) and (ii) assert that the $p_{ij}(t)$'s are probabilities. Property (iii) is a group reversibility property specifying a conservation principle, that "The rate at which group i individuals mix with group j individuals is the same as the rate at which group j individuals mix with group i individuals." Property (iv) says that some populations may become extinct leaving no individuals to mix with. In the above model the C_i's are assumed constant; if however they were allowed to vary, then Property (iv) would also express the fact that if C_i becomes zero, then the mixing rate of individuals in group i is also zero, that is they no longer mix. The set $\{p_{ij}(t)\}$ is also called a mixing/pair-formation matrix.

There is always a trivial solution of the above framework; when all the groups are isolated:

$$p_{ij}(t) = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}.$$

If all the C_i's are positive then there is always a second solution: random or proportionate mixing. In this case $p_{ij}(t)$'s are independent of i, and can be denoted by $\overline{p}_j(t)$, which (using properties (ii) and (iii)) may be written in the form shown in Table 2.

The $\overline{p}_j(t)$'s satisfy all the mixing properties and provide a useful null model in the study of human sexual/social interactions. Despite the fact that humans do not mix at random, proportionate mixing has been used extensively (and successfully!) in addressing, through mathematical models, issues related to the dynamics and management of communicable diseases (see Castillo-Chavez *et al.* 1988, 1989d and references therein). However Sattenspiel (1987a,*b*) using models with social structure, has clearly demonstrated the fundamental role played by nonrandom mixing in disease dynamics. Consequently, nonrandom mixing should no longer be ignored.

We remark that, in general, the time dependence of the mixing probabilities is not direct, but, as seen above, may rely on the time-dependent sizes of each subpopulation. As the size of each subpopulation changes with time so do the mixing probabilities. When we assume that the mixing probabilities change over time as in "preferred" mixing (Anderson *et al.* 1989, 1990; Castillo-Chavez et al. 1989b; Huang 1989; Cooke *et al.* 1991; Huang *et al.* 1992; Nold 1980; Jacquez *et al.* 1988; Hethcote and Yorke 1984) then we implicitly force changes in behavior over time (as the *reserved* proportion for within group mixing remains the same for all time). The study of the effects that time-dependent behavioral changes have over the mixing probabilities is, of course, of undeniable importance. We feel however, that this study should be conducted in a framework that allows the free incorporation of arbitrary (observed and/or postulated) patterns of change. Forms of mixing, like preferred mixing,

while useful and appealing, unnecessarily constrain the dynamics of the mixing subpopulations: Why should the proportion for within group mixing remain the same when, as in the case of AIDS, the disease induced mortality is so high? In terms of our formulation, using preference (describe by the matrix ϕ below), preferred mixing corresponds to the case in which the elements of preference matrix depend on the set $\{\overline{p}_k(t)$: k =1,2,..,N$\}$ in a very specific way. For a further discussion and elaboration of this point see (Blythe *et al.* 1991; Castillo-Chavez and Busenberg 1991; Castillo-Chavez *et al.* 1991).

Clearly, to evaluate the effects of social structure, we need ways of representing, if possible, all forms of time-dependent mixing in transparent and useful forms. We (Busenberg and Castillo-Chavez, 1989, 1991) have determined a formula that represents all forms of mixing as deviations from random or proportionate mixing (for a simple detailed biological description, see Blythe *et al.* 1991). This formula, giving the time-dependent mixing probabilities, will be used as our model for mixing.

To describe this formula for the $p_{ij}(t)$'s, we need some definitions. Let $\overline{p}_j(t)$ denote proportionate or random mixing, and $\phi = \{\phi_{ij}\}$ denote a preference matrix (a measure of the deviation from proportionate mixing). Let $R_i(t)$ (see Table 2) provide a weighted time-dependent measure of the i^{th} deviation, due to the preferences ϕ_{ik}'s, from uniform or homogeneous mixing. We require (as in Busenberg and Castillo-Chavez, 1989, 1991) that $0 \leq R_i(t) \leq 1$ for all i = 1,2,..,N, and that at least one of the $R_i(t)$ is greater than zero. In general, the matrix ϕ is frequency dependent; consequently ϕ depends on the model (in our case on the set of differential equations describing the epidemic) as the relative sizes of the different groups will change with time. The nature of this dependency cannot be given explicitly (except for few special cases such as in preferred mixing) and cannot be arbitrarily selected because the constraints on the R_i's have to be maintained. These constraints imply that each of the expected values of the ϕ_{ik}'s–with respect to the weights $\overline{p}_k(t)$, k = 1,..,N– must lie in the interval [0,1]. This situation suggests the following question: Is there a rich enough class (for modeling purposes) of matrices ϕ that satisfy the required constraints for all possible dynamical models? The

answer is yes, a sufficient condition is that all the ϕ_{ik}'s are constant and satisfy: $0 \leq \phi_{ik} \leq 1$ i, j = 1, 2, 3,.., N. It is in this general setting, which is independent of the choice of dynamical system, that the estimation problem is formulated. However, we first write a formula that describes all mixing solutions $p_{ij}(t)$ using appropriate measures of the deviations from proportionate mixing (see Table 2). The constraint (iii) implies that $\phi_{ij} \equiv \phi_{ji}$, i.e., the ϕ's are symmetric (a rather more complicated relation must be assumed for the two-sex version of this framework, see Castillo-Chavez and Busenberg, 1991).

Remarks: Although the formula for $p_{ij}(t)$ looks very complicated, it is actually quite intuitive (see 42Blythe *et al.*). We note for example that random mixing, which corresponds to no-preference, is described by letting all ϕ_{ik} equal the constant U (that is $\phi_{ik} = U$ for i, k = 1,2,3,...,N). If we substitute these values into the definition of $R_i(t)$, use the condition $0 \leq R_i(t) \leq 1$, and note that not all $R_i(t)$ can be simultaneously equal to zero, then we must have that U satisfies $0 \leq U < 1$. Substituting U into the equation for $p_{ij}(t)$, and performing some algebra one shows that $p_{ij}(t) \equiv \overline{p}_j(t)$ for all time. Hence no preference implies random or proportionate mixing (Figure 1). If on the other hand the ϕ_{ik}'s are chosen to reflect some degree of preference for individuals belonging to the same group, i.e. like-with-like mixing, then the mixing probabilities move away from proportionate mixing (Figure 4). Although the Figures are quite appealing, some caution is in order, especially when we consider the fact that HIV infection will usually prefigure a lethal disease, so that the dynamics and hence the mixing probabilities can be significantly affected by disease-induced mortality in the high risk groups. Further, these Figures only provide us with a snapshot at a *particular* time – we usually cannot deduce the plot of a movie by a single frame! Figures 1-6 provide snapshots, at different times, of two different families of $p_{ij}(t)$'s . Finally, we note that by choosing the ϕ_{ik}'s to be (possibly distinct) constants for all time, we are implicitly assuming that the preferences of individuals do not change over time, or equivalently, that we have formulated the mixing probabilities in terms of the *initial* preferences (or *initial* deviations from random mixing). We will take advantage of our choice of constant ϕ matrix to estimate the ϕ_{ik}'s from a single sample, i.e. a single set of values of $p_{ik}(0)$ data,

which is denoted by the $N \times N$ matrix of constants $\{d_{ij}\}$. To model changes in behavior we will have to model the ϕ_{ik}'s as time dependent functions; this would however require data that are not at present available. If, however, we only want to explore the effects of theoretical behavioral changes, we can accomplish this through the use of a time-dependent preference matrix and time-dependent average rates of partnership change.

4. ESTIMATING THE MATRIX ϕ FROM ONE SAMPLE

Our objective is to calculate a set $\{\phi_{ik}\}$ which minimizes the distance between $\{d_{ik}\}$, the data in the form of an empirical mixing function, and the model from Table 2, at a particular time. As the model equation holds true for all time t, we may choose $t = 0$, i.e. we are fitting $\{p_{ik}(0)\}$. The $\{\phi_{ik}\}$ must be bounded, i.e., $0 \leq \phi_{ik} \leq 1$, and symmetrical $\phi_{ik} = \phi_{ki}$ for all i, j =1, 2, 3,..., N. A reasonable choice for the objective function is

$$S_1(\{\phi_{ij}\}) \equiv \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(d_{ij} - p_{ij}(0))^2, \tag{1}$$

that is, the mean squared residual between data and model (at a fixed time, usually $t = 0$) used in nonlinear regression. However, numerical simulations show that for symmetric $\{\phi_{ij}\}$, not surprisingly, (1) has an infinite number of solutions. The model is underidentified with respect to the data from a single slice of time. There are N constraints on the $p_{ij}(0)$ (from Property (i)) and N(N-1)/2 from Property iv, while the matrix ϕ is only constrained to be symmetric with entries in [0,1]. Thus, the same minimum value of S_1 can be achieved in an infinite number of ways. Even worse, many or all of these may involve ϕ_{ij} that do not lie between 0 and 1, the assumed acceptable region. In mathematical jargon the solutions lie in a surface. This lack of uniqueness does not contradict biological thought as it can be seen from the population genetics literature regarding the relationship between mating

preferences and mating patterns (see Gimelfarb, 1988a,b, and references therein). For example, assortative mating preferences may generate (due to frequency and density dependent effects) random mating patterns. This is also (not obviously) the case for our general mixing matrix $\{p_{ij}(t)\}$. Recently, it has been established (see Palmer *et al.* 1991) that all constant ϕ matrices that lead to random mixing live, in the 2-group case, on a complicated surface in a three-dimensional space.

Of course, part of this problem of non-uniqueness arises because we have an estimate (from data) of $\{p_{ij}(t)\}$ at a single time. If one or more subsequent estimates of $\{p_{ij}(t)\}$ are available then, provided the $\{\overline{p}_j(t)\}$ are also estimated, the objective function S_1 may in principle have a unique minimum. The collection of longitudinal $\{p_{ij}(t)\}$, $\{\overline{p}_j(t)\}$ and $\{C_j(t)\}$ data constitutes a formidable task, as witnessed by the fact that estimates for a single time have not yet been achieved. In this Section, we introduce a technique for making the most of a single "time-slice" of mixing data, which allows us to partially avoid the non-uniqueness and non-boundedness problems described above. The main objective of introducing a method for estimating the matrix ϕ is to formulate the problem, to illustrate potential sources of difficulty, and to instigate further research in this important theoretical and practical problem. We do not wish to imply that the method of this article gives accurate results but rather to illustrate a possible approach.

We do this by introducing a "penalization" factor to the fitting procedure, somewhat in the spirit of those used for the smoothing of spline approximations. Formally, we replace the objective function Eq (1) by the new perturbed objective function

$$S(\{\phi_{ij}\}) \equiv S_1(\{\phi_{ij}\}) + \lambda S_2(\{\phi_{ij}\}), \tag{2}$$

where $S_1(\{\phi_{ij}\})$ is (1) and $S_2(\{\phi_{ij}\})$ is an appropriately chosen penalty function. Unfortunately, there is not a natural choice, and different choices will lead to different solutions. However, we feel that this

problem is too important for us to simply throw up our hands in despair. The choice in this paper (for illustrative purposes) is arbitrary, namely

$$S_2(\{\phi_{ij}\}) \equiv \frac{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(d_{ij}-\phi_{ij})^2}{|\bar{\Phi}|}, \tag{3}$$

in which $\bar{\Phi}$ is the the average of the N^2 ϕ_{ij} values. The absolute value $\bar{\Phi}$ is used because we wish to keep S_2 positive, and negative ϕ_{ij} may enter during intermediate steps in the fitting procedure. This can introduce biases and may even lead to negative $\bar{\Phi}$. The parameter λ is a nonnegative penalization parameter. It is not hard, however, to think of alternative (possibly more appropriate) penalty functions. For example, we may wish to choose the constant ϕ matrix that is "closer" to a ϕ-matrix that gives rise to random mixing. This ϕ matrix could be defined by letting $\phi_{ij} \equiv \bar{\Phi}$ (the average of the N^2 ϕ_{ij} values) for all i, j = 1, 2,..., N., and then using

$$S_2(\{\phi_{ij}\}) \equiv \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\bar{\Phi}-\phi_{ij})^2 .$$

instead of the form used in Eq (3). We use (3), an arbitrary penalty factor, as we only wish to provide a *solution* to the problem of estimating the mixing probabilities. There are of course other theoretical approaches to the definition of a penalty function for the estimation procedure of this article, and likewise there are alternative techniques (requiring some specific assumptions), such as Maximum Likelihood Estimation, that could also be applied. We have began our work in these directions motivated by the fact that a *unique* acceptable solution to the fitting procedure can actually be identified from a *single* sample.

For large λ, S_1 becomes irrelevant to the minimizing process, and we obtain $\phi_{ij} \simeq d_{ij}$, a result which will be unique and properly bounded, but useless for dynamic modeling because the $\{d_{ij}\}$ may be

very far from the $\{p_{ij}\}$. As λ is reduced towards zero, the contribution of S_2 drops accordingly, and minimization of Eq (3) give us a fit of $\{p_{ij}\}$ to $\{d_{ij}\}$ (the lower S_1 the better the fit), with the contribution from S_2 tending to constrain the ϕ, but reducing the quality of the fit. This is an inverse non-linear optimization, on the basis of the equation in Table 2, and subject to the criterion Eq (2), and a s such may not in general have a unique solution for finite λ (different starting values of ϕ might lead to different final fitted values). Empirically, we have found that if the initial guesses for the ϕ are in [0,1], then we usually obtain a unique solution for large λ, but as λ approaches zero, multiple solutions may be found. Our pragmatic approach is to look for the smallest λ such that we do not find multiple solutions, given the prescribed range of initial guesses. We will (somewhat loosely) refer to a solution which arise for λ greater than this critical value as *unique*.

The problem of [0,1] boundedness of the $\{\phi_{ij}\}$ remains, however. The best we can do here is simply to use as the "best fit" the $\{\phi_{ij}\}$ for the smallest λ where *both* the uniqueness (in the above sense) *and* boundedness are not violated. We call this value of $\lambda \equiv \lambda_c$. Fluctuations on the value λ_c will be a function of the values of $\{p_{ij}(t)\}$, $\{\overline{p}_j(t)\}$, and $\{C_j(t)\}$, but for well-posed problems λ_c seems usually to be between 0.0 and 0.1. In Appendix B, we describe the algorithm that we utilized in this estimation procedure.

We stress the fact that the choice of the penalty factor is arbitrary. Our main objective is to provide a *solution* to the problem of estimating the mixing probabilities. We re-emphasize that our research is motivated by the fact that a *unique* acceptable solution to the fitting procedure of our model can actually be obtained from a *single* sample.

5. EXAMPLES

In this section we illustrate our methods for three different values of N, namely N = 2, 3 and 6. We used published "data" that was arbitrarily constructed using (in some cases) some partial information. At present there is *no* data on the matrix $p_{ij}(t)$'s at any particular time for any set of interacting subpopulations, that is, we do not know who is mixing with whom. Crawford *et al.* (1990) have just conducted a survey on sexual behavior at Cornell University, and Castillo-Chavez *et al.* (1991) and Rubin *et al.* (1991) have obtained some estimates (under some restrictive assumptions) of mixing patterns for college-age students. The estimation of these particular mixing matrices is extremely difficult because of (among other factors) the lack of information on size of "external" mixing populations, that is, the lack of knowledge on the size of the subpopulation of non-college sexual partners of those (sexually active) college students that participated in the survey. Because of our use of "pseudo-data", the results in this Section are not intended to be representative of any realistic situation. They are used for illustrative purposes only.

The estimation works with different degree of success in all cases. However, increasing the dimensionality implies, in many realistic situations, smaller values for the mixing probabilities $\{d_{ij}\}$ than would probably be obtained from survey data. Small mixing probabilities place limitations on the applicability of our one-parameter (λ) penalization procedure for estimating the $\{p_{ij}(0)\}$ from a single sample. A more detailed discussion of these and other issues related to our approach is found in the next section.

N = 2

This is the simplest example of a mixing framework, corresponding to a "core group plus other" classification. We use the parameter and "data" shown in Table 3. These data are extracted from (39), the d_{ij} are just arbitrary perturbations from random mixing. The 2x2 is a special, simple case, and the

behavior is straightforward. Figure 7 shows the variation of total S and of S_1 as λ is decreased from $1 \mapsto 0$. We note that S and S_1 here are the respective averages of 5 replicates (a set of 5 randomly chosen initial guesses for the ϕ_{ik}'s were taken). Here S_1 is almost constant, dropping a little as $\lambda \mapsto 0^+$ (i.e. λ approaches 0 while taking only positive values) and total S is almost linear. For large λ, what we see is the minimization of S_2, with S_1 almost constant. This means that we fit ϕ to d directly, and that almost any fit is about as good or bad as any other when we consider the value of S_1. The λ_c is not obvious from Figure 7. We must look at the estimated $\{\phi_{ij}\}$ themselves for this. In this simple case we do not run into the problem of unacceptable ϕ's as $\lambda \mapsto 0^+$, where there is a region in which uniqueness is lost. This region where uniqueness is lost is more sharply defined in Figure 8, where we have plotted one of the ϕ's (ϕ_{11} in fact) against λ. ϕ_{11} decreases with decreasing λ, until we hit $\lambda \simeq \lambda_c$, after which the various *different* ϕ solutions introduce uncertainty. In this case the λ_c can be identified with high degree of accuracy.

N = 3

Again we performed a series of runs, taking 5 replicates at each λ to test for uniqueness, and noting where the ϕ-acceptability was violated. Figure 9 shows the S and S_1 average cases. For large λ (i.e. $\lambda \simeq 1$) they are similar to those of Figure 7 for N = 2, but curves defining S and S_1 as $\lambda \mapsto 0^+$ become more nonlinear in appearance. Again note that even for nonunique selection the S_1 are the same, and obviously as λ gets smaller, S_1 and S approach a common value. By plotting ϕ_{11} against λ (Figure 10) as an indicator, we can see that λ_c must be very small in this case (mainly because the d_{ij} are not uniformly near zero or 1). A simple test to hunt for λ_c consists of looking for values of λ for which $-\epsilon \leq \phi_{ij} \leq 1 + \epsilon$, $0 < \epsilon << 1$. We have observed that ϵ's around 10^{-6} seem to work well. Here we have a well-behaved solution with low λ_c and relatively good fidelity of estimated p's to data.

N = 6

This case is *partially* based upon the artificial test data of Anderson et al. (7). The d_{ij} they used

lie in the region where the problem of unacceptable and non-unique values of ϕ_{ij} (obtained by our algorithm) is more severe (see the discussion in the next section), so we have chosen to use the same $\overline{p}_j(0)$'s, but have pseudo-randomly selected the rest of the required data. The mixing probabilities given by the matrix d $=\{d_{ij}\}$) are "randomly" selected while we require that the entries satisfy the properties (i) and (ii) (they are probabilities and the rows of the matrix d sum to one.)

This case represents a severe test of the ϕ estimation technique because the d's being pseudo-randomly chosen may not reflect viable forms of mixing that would impose some structure over the $\phi_{i,j}$'s. In this case, some of them are very small, which tends to enlarge the range of λ where the values of the ϕ's are unacceptable, hence increasing λ_c. Figure 11 shows how S and S_1 vary with λ. We use the data of Table 5 to illustrate the shortcomings of the method. For large N problems, it becomes difficult to find a λ_c small enough (see Figure 12) that the fit of the estimated values of the p's given by the matrix e $=\{e_{ij}\}$ to the data (matrix d) may be acceptable for dynamic modeling purposes. Until large N techniques are developed and/or more longitudinal data become available, highly aggregated models probably represent the upper limit of modeling. Based on our earlier work (6,14), we have began to work on alternative solutions for large N. For example, we may constrain the ϕ_{ij}'s to a class, say $\phi_{ij} = \phi(|i - j|)$. This would reduce the number of terms to estimate, and provide an "automatic" penalization function. In fact, to guarantee uniqueness, it is clear that for N = 2 a one-parameter ϕ is required while for N = 3 a 3(or less)-parameter ϕ is required. From published data on sexual behavior we may (in particular cases) postulate a realistic parametric ϕ. For example, our analysis (Castillo-Chavez *et al.*, 1991; Rubin *et al.*, 1991) shows that there is a strong like-with-like component in the mixing patterns of college students (not necessarily on their preferences). This pattern may be the result of a very complex or a very simple ϕ and models can help us identify "simple" ϕ's and hence help us formulate testable preference hypothesis. The measurement of preferences may be less difficult from the technical (and unfortunately the political) point(s) of view.

6. DISCUSSION

There are three important comments on the penalization technique which must be borne in mind. First, the quality of the output must be entirely dictated by that of available data. Primarily, this means that there is a limit to how good a fit can be on the basis of just a "one-time" slice of data. In many cases the matrix of estimated values of $p_{ij}(0)$'s, namely $e = \{e_{ij}\}$, which we would wish to use as initial mixing values in a dynamic model (see Appendix A), will be unacceptably far from the original d_{ij}. There is no magical way of getting around this – insufficient data will always wreck models, no matter how beautiful. Our investigations suggest that this problem becomes marked at large values of N, because in many instances unacceptable ϕ_{ij} values occur while λ is still not small enough. For larger N, the constraint of having only one penalization parameter to vary becomes too restrictive, as small d_{ij} (which occur more often for larger values of N) tend to lead to larger values of λ_c in order to remain within the region of acceptable values of the ϕ_{ij}'s. This is a common problem in the biological and social sciences involving the tradeoffs of using "realistic" (large, many parameters) versus tractable (small, few parameters) models. In many instances tractable models are more efficient (Ludwig, 1989; Blythe and Anderson, 1988*b*), and our simulations suggest that $N \simeq 5$ is about the upper limit of groupings for which parameters may be estimated on the basis of one "time-slice," using this penalization method. This has important implications for modelers and social scientists.

The second comment is not unrelated to the first, and concerns testing the penalization technique. Any evaluation requires data, which we do not yet have, so artificial or simulated data must be used. Caution is advised: the obvious temptation is to choose a set of test $\{p_{ij}\}$ which already satisfy the constraints (i-iv), and then estimate the $\{\phi_{ij}\}$. In fact, the use of perfect data to test our algorithm is the worst possible thing to do. The reason for this is intimately related to the problem of the infinite number of solutions to Equation (1), because in the case of perfect data, the $\{p_{ij}(0)\}$ can match *exactly* the $\{d_{ij}\}$, the estimator is operating in a regime where non-uniqueness applies even for rather large λ,

and unacceptable $\{\phi_{ij}\}$ are common. It is thus almost impossible to get a good estimate of $\{p_{ij}(0)\}$.

The third comment is not unrelated to the previous two, and concerns the estimation of the mixing matrices from available data. The number of parameters involved in the estimation of the matrix $\{d_{ij}\}$ increases considerably with N. Hence, from this practical point of view it becomes unrealistic to consider more than six groups (see Crawford *et al.*, 1990). If the objective of developing models for AIDS is to produce some possibly useful results, we must include these data-oriented considerations in our theoretical studies.

We conclude with some comments regarding the approach of this article. There are two features clearly and sharply illustrating the limitations of the technique, which are useful in deciding its applicability in any given situation. We would like to reiterate that these problems mainly occur because we are trying to estimate $\{\phi_{ij}\}$ from data at a single "time-slice." Two $\{p_{ij}(t)\}$ estimates at different (but fairly close) times, even if rather poor, will be of greater utility than one very high quality sample, precisely because the $\{\phi_{ij}\}$ can take so many possible values.

7. CONCLUSIONS

We are at present developing alternative techniques to get $\{\phi_{ij}\}$ estimates from one-sample data, that is, from a mixing matrix $\{p_{ij}(0)\}$, and are also constructing the most robust surface-fitting schemes necessary to get a good estimate from the multi time-slice data. We are also using the two-sex mixing models (Castillo-Chavez *et al.* 1989*b*; Castillo-Chavez and Busenberg, 1991), for which the corresponding $\{\phi_{ij}\}$, is not symmetric to test this method for heterosexual populations.

We note, however, the existence of at least two estimation problems, one dealing with the estimation of the matrix $\{p_{ij}(0)\}$ from data and the other with the that of the corresponding ϕ-matrix. The estimation of the mixing matrix $\{p_{ij}(0)\}$ demands knowledge of the sizes of the mixing

subpopulations and, when such knowledge exists, it is done on the unstated assumption of the existence of a *closed* mixing network. Unfortunately, there are no realistic closed networks. Furthermore, studies of college-age mixing subpopulations reveal the existence of networks for which the internal mixing accounts for only 50% of the contacts (see Crawford *et al.* 1991, Rubin *et al.* 1991). Given the relevance of these problems to the issues raised in this manuscript, we conclude this paper with an outline of *preliminary* alternative approaches to some of the estimation problems outlined.

If we assume that the affinities are time independent then the main equation in Table 2 (or Equation 4 below), as discussed earlier, provides a very large class of mixing/pair-formation models from which we can estimate the affinities from data on the mixing probabilities at a single-time slice. We now explicitly state various approaches that we have begun to utilize in order to estimate the time-dependent contact structure of a population.

I. A maximum likelihood estimation approach: The general model is

$$p_{ij}(0) = \bar{p}_j(0)\left(\frac{R_i R_j}{\sum_k \bar{p}_k R_k} + \varphi_{ij}\right), \tag{4}$$

where p_{ij} is the true probability of i-with-j mixing. We observe d_{ij}, where $E[d_{ij}] = p_{ij}$ or, equivalently, we model

$$d_{ij} = \bar{p}_j\left(\frac{R_i R_j}{\sum_k \bar{p}_k R_k} + \varphi_{ij}\right) + \epsilon_{ij} \quad ,$$

where ϵ_{ij} is error, with $E[\epsilon_{ij}] = 0$.

Since p_{ij} is a probability, it is bounded between 0 and 1. Such bounded functions are often difficult to handle statistically, so a transformation is done to "unbound" the range. Common transformations are

$$\log(p_{ij}) \qquad \text{or} \qquad \log\left(\frac{p_{ij}}{1-p_{ij}}\right) \equiv \text{logit}(p_{ij}) \quad .$$

For example, we might try to fit $\log(p_{ij})$, and specify

$$\log (d_{ij}) = \log \bar{p}_j + \log\left[\frac{R_i R_j}{\sum \bar{p}_k R_k} + \varphi_{ij}\right] + \epsilon_{ij} \quad ,$$

where $\epsilon_{ij} \sim$ normal $(0, \sigma^2)$, all independent. This would lead to a likelihood function

$$L = \prod_{ij} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\tfrac{1}{2}\sigma^2\left[\log (d_{ij}) - \log \bar{p}_j - \log\left(\frac{R_i R_j}{\sum \bar{p}_k R_k} + \varphi_{ij}\right)\right]^2\right\} \quad .$$

Now, given that we know $\bar{p}_j$, the only unknowns in L are φ_{ij} and σ^2. Maximization of L over these parameters would give us Maximum Likelihood Estimates. Other variations of this estimation approach include using logit instead of log, or model the p_{ij} directly without a transformation, using a binomial or multinomial model.

II. Parametric φ_{ij}: The φ_{ij} used in either I or in the Least Square Method, used in this manuscript, can be members of a parametric family. In particular, we could have

$$\varphi_{ij} = \begin{cases} a + b & i=j \\ b & i \neq j \end{cases} = b + a\delta_{ij} \quad , \tag{5}$$

where $a + b \leq 1$; $a,b \geq 0$. This class would make optimization over φ_{ij} simple, as we would only have a two parameter class.

Another model is

$$\varphi_{ij} = \varphi(i-j) \quad , \tag{6}$$

which is a generalization of (5) (*i.e.* (5) is a submodel of (6). A special case of (6), which might be reasonable to work with, is

$$\varphi_{ij} = \begin{cases} a + b & i=j \\ bc^{|i-j|} & i \neq j \end{cases} \quad . \tag{7}$$

This is a three-parameter class, of which (2) is a submodel.

In general, we can write

$$\varphi_{ij} = h_{ij}(\underline{a})$$

where h_{ij} is a known function, and $\underline{a} = (a_1, \cdots, a_k)$ is a vector of parameters. We then optimize over these parameters.

As a last example, we could use a logit model for φ_{ij} alone. For example, we could write

$$\text{logit}(\varphi_{ij}) = \log\left(\frac{\varphi_{ij}}{1-\varphi_{ij}}\right) = \alpha + \beta' X,$$

where α and β are parameter vectors, and X is a vector of covariates (age, race, etc.). We then have

$$\varphi_{ij} = \frac{e^{\alpha + \beta' X}}{1 + e^{\alpha + \beta' X}},$$

and we would optimize I or (1) over α and β.

III. Hierarchical (Empirical Bayes) Models: We will discuss a somewhat *ad hoc* scheme for empirical Bayes estimation. The scheme can be made more formal once the exact type of model is determined. What follows is an all-purpose general approach.

Consider two models for φ_{ij}, one of which is a submodel of the other, referred to as Full model and Submodel. These can be any two models, but for convenience think of

$$\varphi_{ij}^F = b + a\delta_{ij} \qquad \text{(Full model)}$$

$$\varphi_{ij}^S = b \qquad \text{(Submodel)}$$

The submodel (proportionate mixing) is a special case of the full model (a=0).

For each model we can compute the residual sum of squares, as in (1), as

$$RSS_F = \min_{\substack{a,b \\ a+b \le 1}} \left\{ \sum_{i,j} \left(d_{ij} - \bar{P}_j \left[\frac{R_i R_j}{\sum \bar{P}_k R_k} + \varphi_{ij}^F \right] \right)^2 \right\}$$

$$RSS_S = \min_{b} \left\{ \sum_{i,j} \left(d_{ij} - \bar{P}_j \left[\frac{R_i R_j}{\sum \bar{P}_k R_k} + \varphi_{ij}^S \right] \right)^2 \right\} .$$

Note that $RSS_F < RSS_S$, as the minimization is over a larger set. This is an important consequence of one model being a submodel of the other. Now define

$$T \equiv \frac{RSS_S - RSS_F}{RSS_F},$$

a measure of the goodness of the submodel with respect to the full model. Small values of T (near 0) happen if $RSS_S \approx RSS_F$, and support the submodel. Large values of T will happen if $RSS_S << RSS_F$, and support the full model. (T is actually proportional to an F-statistic--some assumptions are required, and a significance test of the submodel can be done.)

Now calculate the combined estimate (empirical Bayes estimate) of φ_{ij} as follows: We have estimates $\hat{\varphi}^F_{ij}$ and $\hat{\varphi}^S_{ij}$ from the two separate minimizations above. Form the convex combination

$$\hat{\varphi}^{EB}_{ij} = \left(\frac{T}{1+T}\right)\hat{\varphi}^F_{ij} + \left(1 - \frac{T}{1+T}\right)\hat{\varphi}^S_{ij}\,.$$

This is an empirical Bayes estimate of φ_{ij}. Small values of T support the submodel, and pull $\hat{\varphi}^{EB}_{ij} \rightarrow \hat{\varphi}^S_{ij}$. Large values of T pull $\hat{\varphi}^{EB}_{ij} \rightarrow \hat{\varphi}^F_{ij}$. Under reasonable assumptions, $\hat{\varphi}^{EB}$ is the superior estimate.

If one wishes to compute explicit examples of mixing matrices from data on mixing, one needs to estimate the sizes of the mixing subpopulations. Knowledge of these matrices over a period of time is essential to any type of long-term forecasting. Because our purposes are and data are limited, we do not need to use sophisticated approaches in the construction of these matrices. Capture-recapture methodology can be applied to survey data to estimate the number of different sexual partners from each of several groups that an individual has had in a fixed period of time, or to estimate the size of the population having sexual contact with members of a given group. Thus, one can apply this methodology to survey data to estimate the size of the population at risk for a sexually transmitted disease. Using data from our survey conducted at Cornell University in 1989 (see Crawford *et al.* 1991), we have successfully used capture-recapture estimators to provide estimates of the size of the population that has sexual contact with Cornell undergraduates but are not Cornell students (see Castillo-Chavez *et al.* 1991; Rubin *et al.* 1991). ***We have used these estimates and our one- and two-sex mixing framework to construct explicit mixing matrices.*** In our situation, prior to sampling, the population contains both marked and unmarked individuals: contacts (i.e., sexual partners) are either Cornell students or not. We can only access information about Cornell and non-Cornell partners from the Cornell students surveyed. The students surveyed play the role of observers in capture-recapture bird studies in which "recapture" is done by sighting (for more details see Castillo-Chavez *et al.* 1991; Rubin *et al.* 1991).

Capture-recapture estimators are design-based rather than model-based; they do not rely on a probabilistic model. Capture-recapture population estimates can provide an independent benchmark against which to compare estimates based on different probabilistic models. Mixing matrices are finally

constructed by forcing the data and the estimated parameters including the sizes of the mixing subpopulations to satisfy the mixing axioms (i) – (iv).

In this last section we have provided a very rough outline of a program for the estimation of parameters in mixing/pair formation models (other approaches are of course possible, see for example Pugliese 1991). Development of techniques and novel approaches for the validation of STD's and HIV/AIDS models is a matter of considerable importance in the era of AIDS. We conclude by stressing again that the outlined provided above has as its main objective, the instigation of further work in this important area of research.

ACKNOWLEDGMENTS

This research has been partially supported by NSF grant DMS-8906580, NIAID Grant R01 A129178-01, 02, and Hatch project grant NYC 151-409, USDA to CC-C. SPB's research has also been partially supported by funds from the Office the Dean of the College of Agriculture and Life Sciences at Cornell University and the Mathematics Science Institute. We thank K. Cooke, S.-F. Shyu, and an anonymous reviewer, for their valuable comments.

References

Anderson, R.M. (1988) The role of mathematical models in the study of HIV transmission and the epidemiology of AIDS. *J. AIDS* **1**: 241-256.

Anderson, R.M., and May, R.M. (1988) Epidemiological parameters of HIV transmission. *Nature* **333**: 514-519.

Anderson, R.M., Gupta, S., and Ng, W. (1990) The significance of sexual partner contact networks for the transmission dynamics of HIV. *J. of AIDS* **3**: 417-429.

Anderson, R.M., Blythe, S.P., Gupta, S., and Konnings, E. (1989) The transmission dynamics of the human immunodeficiency virus type 1 in the male homosexual community in the United Kingdom: the influence of changes in sexual behavior. *Phil Trans R Soc London B* **325**: 45-89.

Blythe, S.P., and Anderson, R.M. (1988*a*) Distributed incubation and infectious periods in models of transmission dynamics of human immunodeficiency virus (HIV). *IMA J. Math. Appl. Med. & Biol.* **5**: 1-19.

Blythe, S.P., and Anderson, R.M. (1988*b*) Variable infectiousness in HIV transmission models. *IMA J. Ma th. Appl. Med. & Biol.* **5**: 181-200.

Blythe, S.P., and Castillo-Chavez, C. (1989) Like-with-like preference and sexual mixing models. *Math. Biosci.* **96**: 221-238.

Blythe, S.P., and Castillo-Chavez, C. (1990*a*) Scaling of sexual activity. *Nature* **344**: 202.

Blythe, S.P., Castillo-Chavez, C., Palmer, J.S., and Cheng M.A (1991)Towards a unified theory of mixing and pair-formation. *Math. Biosci.* (in press).

Brookmeyer, R., and Gail, M.H. (1988) A method of obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *J. Am. Stat. Assoc.* **83**: 301-308.

Busenberg, S., and Castillo-Chavez, C. (1989) Interaction, pair formation and force of infection terms in sexually transmitted diseases. *Lect. Notes Biomath.* **83**: 289-300.

Busenberg S., and Castillo-Chavez C. (1991) A general solution of the problem of mixing subpopulations, and its application to risk- and age- structured epidemic models for the spread of AIDS. *IMA J. of Mathematics Applied in Med. and Biol.* 8: 1-29

Castillo-Chavez, C. (1989) *Mathematical and Statistical Approaches to AIDS Epidemiology* (ed., C. Castillo-Chavez), Lecture Notes in Biomathematics, No. 83, Springer Verlag, Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong.

Castillo-Chavez, C., and Blythe, S.P. (1989) Mixing framework for social/sexual behavior. *Lect. Notes Biomath.* 83: 275-284.

Castillo-Chavez, C. and Busenberg, S. (1991) On the solution of the two-sex problem. In: *Proceedings of the Interantional Conference on Differential Equations and Applications to Biology and Population Dynamics* (S. Busenberg and M. Martelli, eds.), Lecture Notes in Biomathematics, Springer-Verlag (in press).

Castillo-Chavez, C., Busenberg, S. and Gerow, K. (1991) Pair formation in structured populations. In: *Differential Equations with Applications in Biology, Physics and Engineering* (J. Goldstein, F. Kappel, W. Schappacher, eds.), Marcel Dekker, New York, pp. 47-65.

Castillo-Chavez, C., Cooke, K., Huang, W., and Levin, S.A. (1989a) On the role of long incubation periods in the dynamics of acquired immunodeficiency syndrome (AIDS), Part 1. Single population models. J. *Math. Biol.* 27: 373-398.

Castillo-Chavez, C., Cooke, K., Huang, W., and Levin, S.A. (1989b) On the role of long incubation periods in the dynamics of acquired immunodeficiency syndrome (AIDS), Part 2. Multiple group models. *Lect. Notes Biomath.* 83: 200-217.

Castillo-Chavez, C., Cooke, K., Huang, W., and Levin, S.A. (1989c) Results on the dynamics for models for the sexual transmission of the human immunodeficiency virus. *Appl. Math. Lett.* 2: 327-331.

Castillo-Chavez, C., Shyu, S.-F., Rubin, G., and Umbach, D. (1991) On the estimation problem of mixing/pair formation matrices with applications to models for sexually transmitted diseases (manuscript).

Castillo-Chavez, C., Hethcote, H.W., Andreasen, V., Levin, S.A., and Liu, W.M. (1988) Cross-immunity in the dynamics of homogeneous and heterogeneous populations. pp 303-316 in (L. Gross, T.G. Hallam, and S.A. Levin, eds.) *Mathematical Ecology.* Proceedings, Autumn Course Research Seminars, Trieste 1986. World Scientific Publishing Co., Singapore.

Castillo-Chavez, C., Hethcote, H.W., Andreasen, V., Levin, S.A., and Liu WM. (1989d) Epidemiological models with age structure, proportionate mixing, and cross-immunity. *J. Math. Biol.* **27**: 233-258.

Cooke, K. L., Allers, D. A., and Castillo-Chavez, C. (1991) Mixing patterns models of AIDS. In: *Proceedings of the 2nd International Conference on Differential Equations and its Applications* (Ovide Arino, D. Axelro and M. Kimmel, eds.), Rutgers 1989. pp. 297-309.

Cox, D.R., and Medley, G.F. (1989) A process of events with notification delay and the forecasting of AIDS. *Phil Trans R Soc London B* **325**: 135-145.

Crawford, C. M., Schwager, S. J., and Castillo-Chavez, C. (1990) A methodology for asking sensitive questions among college undergraduates. Biometrics Unit Tech. Report BU-1105-M, Cornell University, Ithaca, New York

Day, N.E., Gore, S.M., McGee, M.A., and South M. (1989) Predictions of the AIDS epidemic in U.K.: the use of the back calculation method. *Phil Trans R Soc London B* **325**: 123-134.

Gimelfarb, A. (1988*a*) Processes of pair formation leading to assortative mating in biological populations: encounter mating model. *Amer. Natur.* **131**, 6: 865-884.

Gimelfarb, A. (1988*b*) Processes of pair formation leading to assortative mating in biological populations: dynamic interaction model. *Theor. Pop. Biol.* **34**: 1-23.

Gupta, S., Anderson, R.M., and May, R.M. (1989) Network of sexual contacts: implications for the pattern of spread of HIV. *AIDS* **3**: 1-11.

Isham, V. (1989) Estimation of the incidence of HIV infection. *Phil Trans R Soc London B* **325**: 113-121.

Jacquez, J.A., Simon, C.P., Koopman, J., Sattenspiel, L., and Perry, T. (1989) Modeling and analyzing HIV transmission: the effects of contact patterns. *Math. Bios.* **92**: 119-199.

Hethcote, H.W., and Yorke, J.A. (1984) *Gonorrhea transmission dynamics and control*, Lecture Notes in Biomathematics 56, Springer-Verlag, Berlin Heidelberg New York Tokyo.

Hethcote, H.W., Van Ark, J.W., Karon, J.M., and Longini Jr., I.M. A simulation of HIV and AIDS in homosexual men in San Francisco. Unpublished manuscript.

Huang, W. (1989) *Studies in differential equations and applications.* Ph.D. dissertation, The Claremont Graduate School, Claremont CA.

Huang, W., Cooke, K., and Castillo-Chavez, C. (1991) Stability and befurcation for a multiple group model for the dynamics of HIV/AIDS transmission. *SIAM J. of Applied Math.* (accepted.)

Karon, J.M., Dondero, T.J., and Curran, J.W. (1988) The projected incidence of AIDS and estimated prevalence of HIV infection in the United States. *J of AIDS* **1**: 542-550.

Karon, J.M., Dondero, T.J., and Curran, J.W. (1989) Predicting AIDS incidence by extrapolating from recent trends. *Lect. Notes Biomath.* **83**: 58-88.

Ludwig, D. (1989) Small models are beautiful: efficient estimators are even more beautiful. *Lect. Notes Biomath.* **81**: 274-281.

Ludwig, D., and Walters, C. (1985) Are age structured models appropriate for catch-effort data? *Can. J. Fish. Aquat. Sci.* **40**: 559-569

Nold, A. (1980) Heterogeneity in disease transmission modelling. *Math. Biosci.* **52**: 227-250.

Palmer, J.S.,, Castillo-Chavez C., and Blythe S.P. (1991) State-dependent mixing and state-dependent contact rates in epidemiological models. Biometrics Unit Technical Report (BU-1122-M).

Rubin, G., Umbauch, D., Shyu, S-F., and Castillo-Chavez, C. (1991). Application of capture-recapture methodology to estimation of size of population at risk of AIDS and/or other sexually-transmitted diseases. Biometrics Unit Tech. Report BU-1112-M, Cornell University, Ithaca, New York.

Sattenspiel. L. (1987*a*) Population structure and the spread of disease. *Human Biology.* 59: 411-438.

Sattenspiel, L. (1987*b*) Epidemics in nonrandomly mixing populations: a simulation. *American Journal of Physical Anthropology.* 73: 251-265.

Sattenspiel, L., and Castillo-Chavez, C. (1991) Environmental context, social interactions, and the spread of HIV. *American J. of Human Biology* 2, 4 (in press).

Schwager, S.J., Castillo-Chavez, C., and Hethcote, H. (1989) Statistical and mathematical approaches in HIV/AIDS modelling: a review. *Lect. Notes Biomath.* 83: 2-35.

Thieme, H.R. and Castillo-Chavez, C. (1989) On the role of variable infectivity in the dynamics of the human immunodeficiency virus. *Lect. Notes Biomath.* 83: 157-176.

Thieme H. R. and C. Castillo-Chavez, *How may infection-age dependent infectivity affect the dynamics of HIV/AIDS*, 1991, Biometrics Unit Tech. Report BU-1102-M, Cornell University, Ithaca, New York

Appendix A

Basic Transmission Model for HIV-Dynamics

The model described in the text intentionally omits several important factors (epidemiological, demographical, etc.) because our main objective is to address the general estimation problem associated with the problem of mixing. The model may be written

$$\frac{dS_i(t)}{dt} = \Lambda_i - B_i(t) - \mu S_i(t) , \tag{A1}$$

$$\frac{dI_i(t)}{dt} = B_i(t) - (\mu + \alpha_i) I_i(t) , \tag{A2}$$

$$\frac{dA_i(t)}{dt} = \alpha_i I_i(t) - m_i A_i(t), \quad i=1,2,\cdots,N . \tag{A3}$$

This model assumes constant removal rates from the infective classes into the AIDS classes. This assumption is certainly unrealistic as it implies a negative exponential incubation period distribution for each group. For more realistic incubation period distributions see (Blythe and Anderson, 1988; 24Castillo-Chavez *et al.* 1989; Thieme and Castillo-Chavez, 1989, 1990; 42Blythe *et al.*). The expression for the i^{th} incidence rate $B_i(t)$ is described in Table 1.

Appendix B

Algorithm

We may now specify exactly how to get our best estimate for $\{p_{ij}(0)\}$ which we would be used in a dynamic mathematical model such as the model described in Appendix A. The procedure may be implemented fairly painlessly using a standard statistics or data analysis package such as GAUSS, SYSTAT, SAS, etc. We used GAUSS. For clarity we describe the simplest version. Features of packages like GAUSS can easily help us to simplify this algorithm through the direct handling of matrices.

- N = the number of groups
- $\overline{p} = \overline{p}_j(t)$ = 1-dimensional array of N proportionate mixing fractions
- $d = d_{ij}$ = 2-dimensional array of NxN observed mixing fractions data
- $\phi = \phi_{ij}$ = 2-dimensional array of NxN parameters to be estimated
- Chooose a value for the penalization parameters λ
- $\hat{\phi} = \hat{\phi}_{ij}$ = 2-dimensional array of NxN initial guesses for $\{\phi_{ij}\}$ required by minimization algorithm.
- Minimization. In Gauss, this requires the single statement
  $\{\phi, S_{min}, dS, H\}$ = OPTMUM ($\hat{\phi}$, &<procedure name>).

In the last step we input the array $\hat{\phi}$ and &<procedure name>, a pointer to a GAUSS procedure written to calculate $S = S_1 + \lambda S_2$. The output is ϕ, the matrix of ϕ_{ij} values; S_{min}, the minimum value of S (as in Equation 2), ; dS, the local gradient of S at S_{min}, which should be precisely zero at a local minimum; and H, the "Hessian" an NxN matrix of covariances of the distributions of the matrix of estimated $\{\phi_{ij}\}$.

- Output e, (the matrix of $\{e_{ij}\}$, the estimated $\{p_{ij}(0)\}$), ϕ, S_1, S_2, and selected summary statistics, at the minimum.

We found that in many cases, the fairly standard optimization design of starting with the method of steepest descent, and then moving to a more efficient algorithm when convergence is steady but slow worked well. Near λ_c, many iterations may be necessary, and sometimes only the steepest descent method guaranteed convergence.

Table 1

Formula for $B_i(t)$–the i^{th}-incidence rate.

$C_iS_i(t)$ ≡ Susceptible partnerships from group i

β_j ≡ Probability of transmission per infected group j-partner

$\frac{I_j(t)}{T_j(t)}$ ≡ Proportion of infected people in group j

$\beta_j \frac{I_j(t)}{T_j(t)}$ ≡ Proportion of infected people in group j

capable of transmitting the disease

$p_{ij}(t)$ ≡ probability of choosing a partner from group j:

this is how a "typical" group i individual mixes

with group j individuals

Therefore, if j = 1, 2,...,N, (i.e. N groups) then the number of new cases of infection per unit time in group i is given by

$$B_i(t) \equiv C_iS_i(t) \sum_{j=1}^{N} p_{ij}(t)\, \beta_j \frac{I_j(t)}{T_j(t)}.$$

Table 2

One-sex framework

Proportionate or random mixing:

$p_{ij}(t)$ is independent of i and is denoted by $\overline{p}_j(t)$

$$\overline{p}_j(t) \equiv \frac{C_j T_j(t)}{\sum_{k=1}^{N} C_k T_k(t)}$$

General solution: one-sex mixing/pair-formation problem:

$$p_{ij}(t) \equiv \overline{p}_j(t) \left[\frac{R_i(t)\, R_j(t)}{\sum_{k=1}^{N} \overline{p}_k(t) R_k(t)} + \phi_{ij} \right]$$

Definitions:

$\phi = \{\phi_{ij}\} \equiv$ Time independent initial preference matrix, measures the deviation in preference from random mixing.

$\phi_{ij} \equiv \phi_{ji}$, i.e. the ϕ-matrix is symmetric. This is a consequence of the required properties of the mixing probabilities.

$R_i(t) \equiv 1 - \sum_{k=1}^{N} \overline{p}_k(t) \phi_{ik}$: a weighted time-dependent measure of deviation from uniform mixing

Table 3

Data for Example 1: case N =2

(Modified from Hethcote and Yorke (1984))

Group	C_j	T_j	$\bar{p}_j$
1	1 per month	50, 400	0.4736
2	10 per month	5,600	0.52634

$$d = \begin{bmatrix} .51 & .49 \\ .53 & .47 \end{bmatrix}$$

Table 4

Data for Example 2: case N =3

$$\overline{p} = [0.6 \quad 0.3 \quad 0.1]$$

$$d = \begin{bmatrix} .5 & .3 & .2 \\ .1 & .4 & .5 \\ .6 & .2 & .2 \end{bmatrix}$$

Table 5

Data for Example 3: case N =6

(The C_i and the $T_i(0)$ are from Anderson et al. (1990))
(The elements of d are chosen in a pseudorandom form, see the text for details)

$\bar{p}$ = [0.02811 0.055087 0.07972 0.12573 0.34659 0.36898])

C_i = [0.45000 3.200000 7.02000 13.8400 43.6000 81.2300]

T_i = [0.55000 0.140000 0.10000 0.0800 0.076000 0.04000]

$$d = \begin{bmatrix} 3.96\times10^{-3} & 2.88\times10^{-1} & 6.80\times10^{-1} & 1.33\times10^{-2} & 8.80\times10^{-3} & 6.23\times10^{-3} \\ 1.67\times10^{-1} & 4.16\times10^{-2} & 3.08\times10^{-1} & 4.15\times10^{-2} & 4.86\times10^{-2} & 1.87\times10^{-2} \\ 8.07\times10^{-1} & 7.30\times10^{-2} & 2.66\times10^{-2} & 7.80\times10^{-2} & 7.41\times10^{-3} & 6.75\times10^{-3} \\ 5.93\times10^{-2} & 5.93\times10^{-1} & 2.49\times10^{-1} & 6.05\times10^{-2} & 9.97\times10^{-3} & 2.90\times10^{-2} \\ 7.09\times10^{-1} & 2.53\times10^{-1} & 3.11\times10^{-2} & 5.61\times10^{-3} & 8.29\times10^{-4} & 5.30\times10^{-4} \\ 1.75\times10^{-1} & 3.68\times10^{-1} & 5.92\times10^{-2} & 2.85\times10^{-1} & 5.81\times10^{-2} & 5.49\times10^{-2} \end{bmatrix}$$

Legends of Figures

Figure 1.

Proportionate mixing $\overline{p}_j(0)$: $\phi_{ij} = 0$ for i, j = 1, 2, 3, 4, 5; five groups.

We note that the $T_i(t)$ used here and in Figures 2-6 were obtained as solutions of an epidemic model with variable population size for the spread of gonorrhea (our sole objective is to illustrate the time dependence of the mixing matrix). We note that similar graphs can be obtained using models for the spread of HIV/AIDS.

Figure 2.

N = 5. Proportionate mixing $\overline{p}_j(50)$: $\phi_{ij} = 0$ for i, j = 1, 2, 3, 4, 5.

Figure 3.

N = 5. Proportionate mixing $\overline{p}_j(100)$: $\phi_{ij} = 0$ for i, j = 1, 2, 3, 4, 5.

Figure 4.

N = 5. Like-with-like mixing $p_{ij}(0)$: $\phi_{ij} = 0$ for i, j = 1, 2, 3, 4, 5 ($i \neq j$); $\phi_{ii} = 1$ for i= 1, 2, 3, 4, 5

Figure 5.

N = 5. Like-with-like mixing $p_{ij}(50)$: $\phi_{ij} = 0$ for i, j = 1, 2, 3, 4, 5 ($i \neq j$); $\phi_{ii} = 1$ for i= 1, 2, 3, 4, 5

Figure 6.

N = 5. Like-with-like mixing $p_{ij}(100)$: $\phi_{ij} = 0$ for i, j = 1, 2, 3, 4, 5 ($i \neq j$); $\phi_{ii} = 1$ for i= 1, 2, 3, 4, 5

Figure 7.

N = 2. Plot of minimum values of S and S_1 against the penalization parameter λ. Data from Table 3.

Figure 8.

N = 2. Plot of mean value (5 replicates) of ϕ_{11} against λ. Data from Table 3.

Figure 9.

N = 3. Plot of minimum values of S and S_1 against the penalization parameter λ. Data from Table 4.

Figure 10.

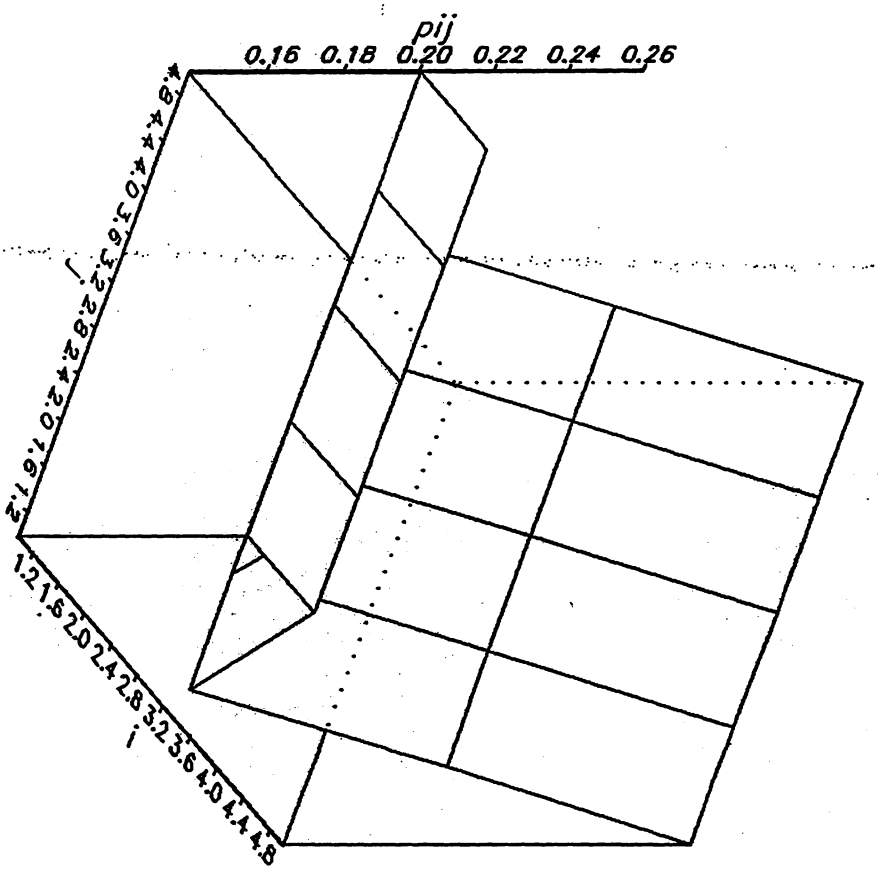
N = 3. Plot of mean value (5 replicates) of ϕ_{11} against λ. Data from Table 4.

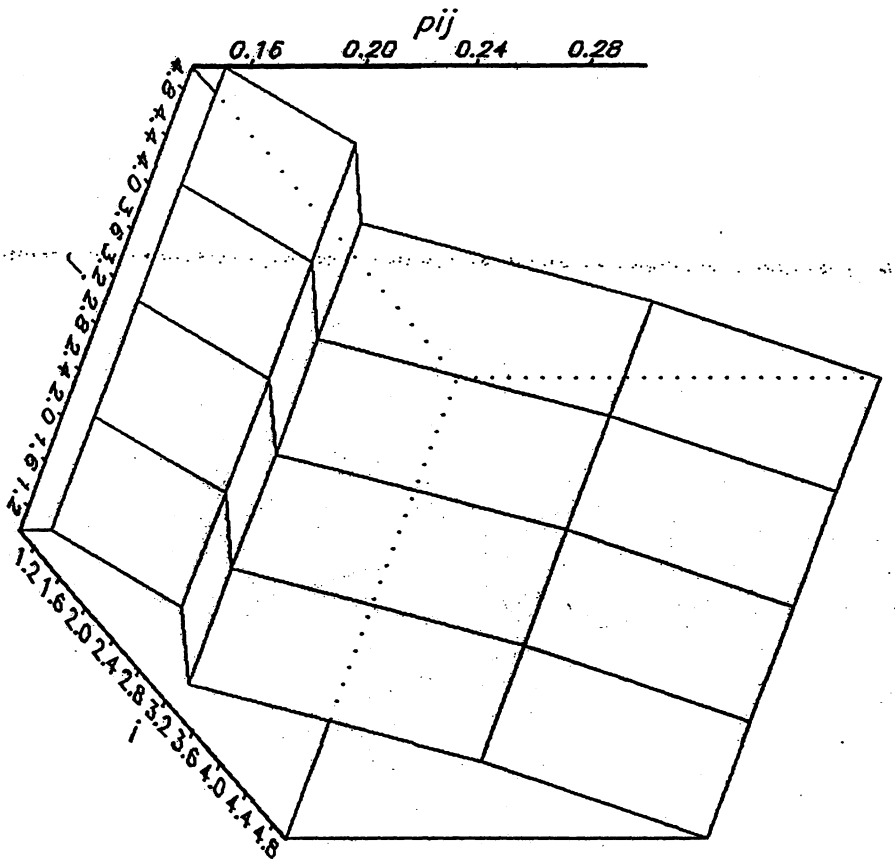
Figure 11.

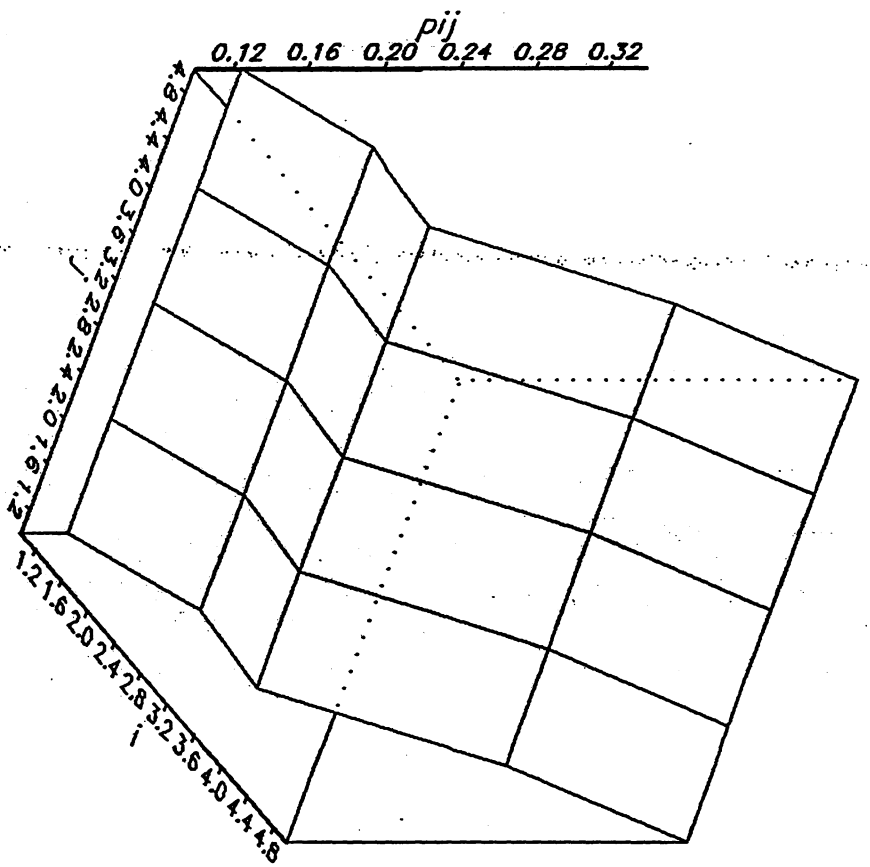
N = 6. Plot of minimum values of S and S_1 against the penalization parameter λ. Data from Table 5.

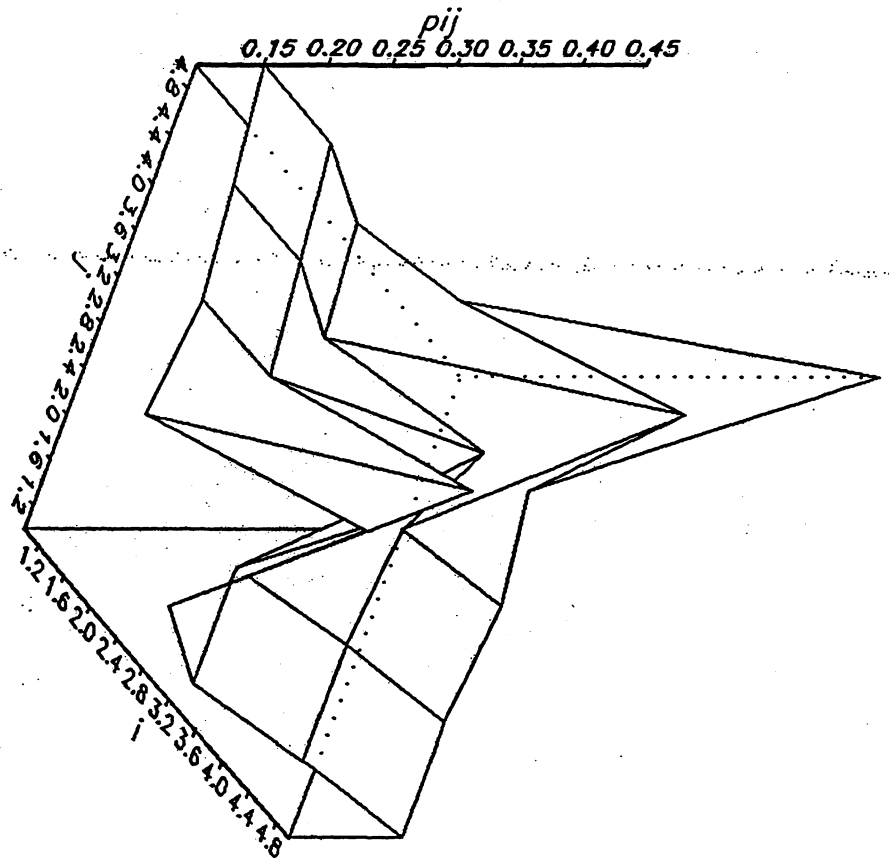
Figure 12.

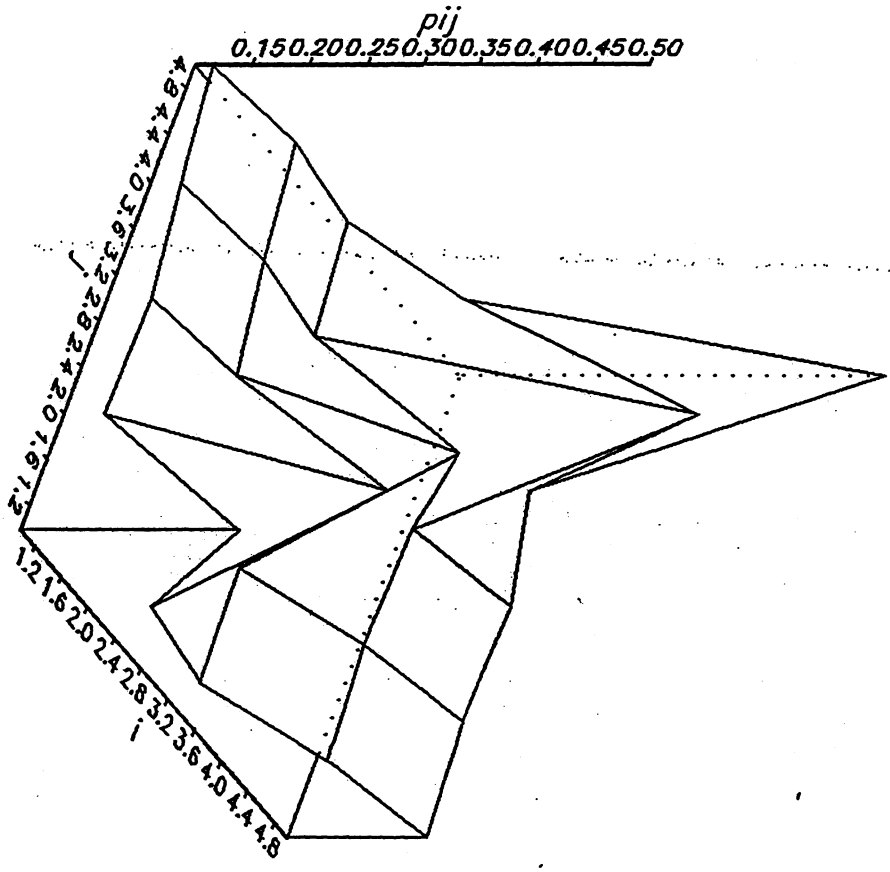
N = 6. Plot of mean value (5 replicates) of ϕ_{63} against λ. Data from Table 5.

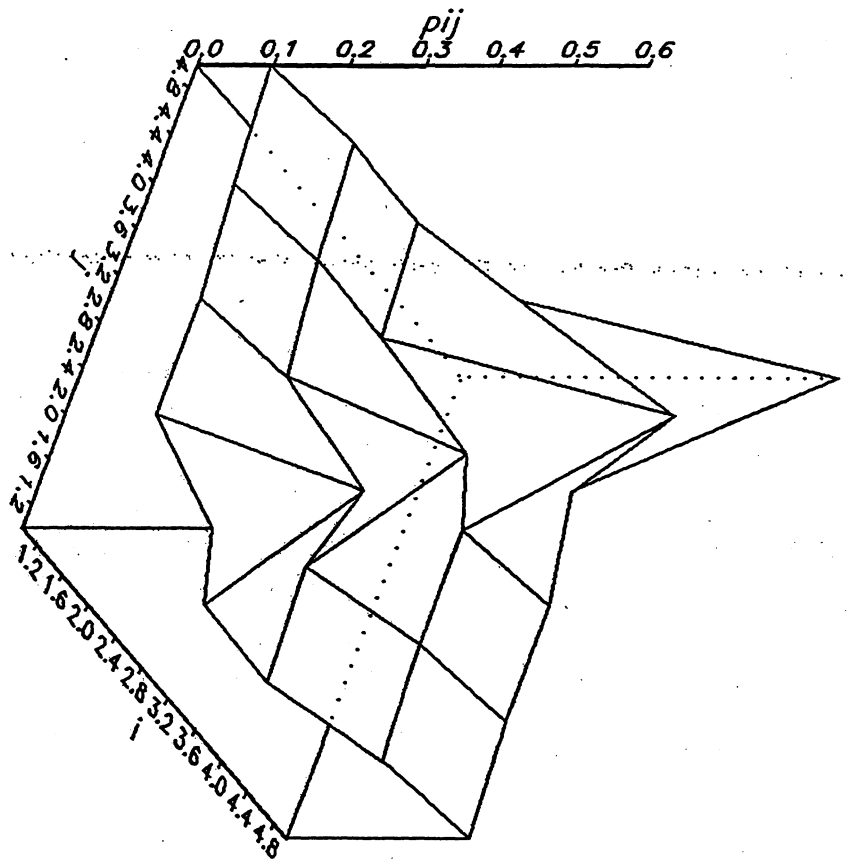

pij
0.16 0.18 0.20 0.22 0.24 0.26
j
4.8 4.4 4.0 3.6 3.2 2.8 2.4 2.0 1.6 1.2
i
1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8

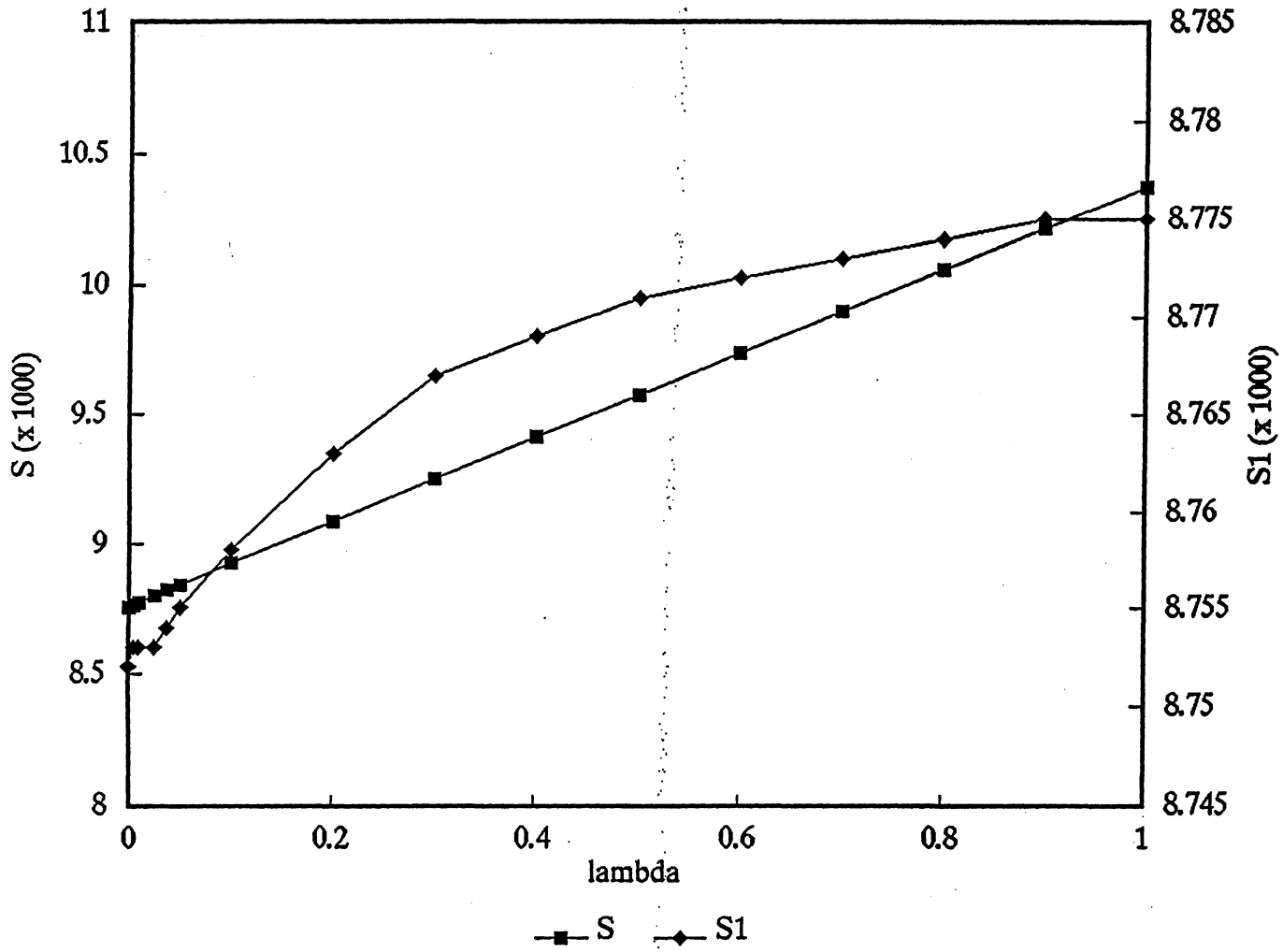

pij
0.16
0.20
0.24
0.28
j
4.8 4.4 4.0 3.6 3.2 2.8 2.4 2.0 1.6 1.2
i
1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8

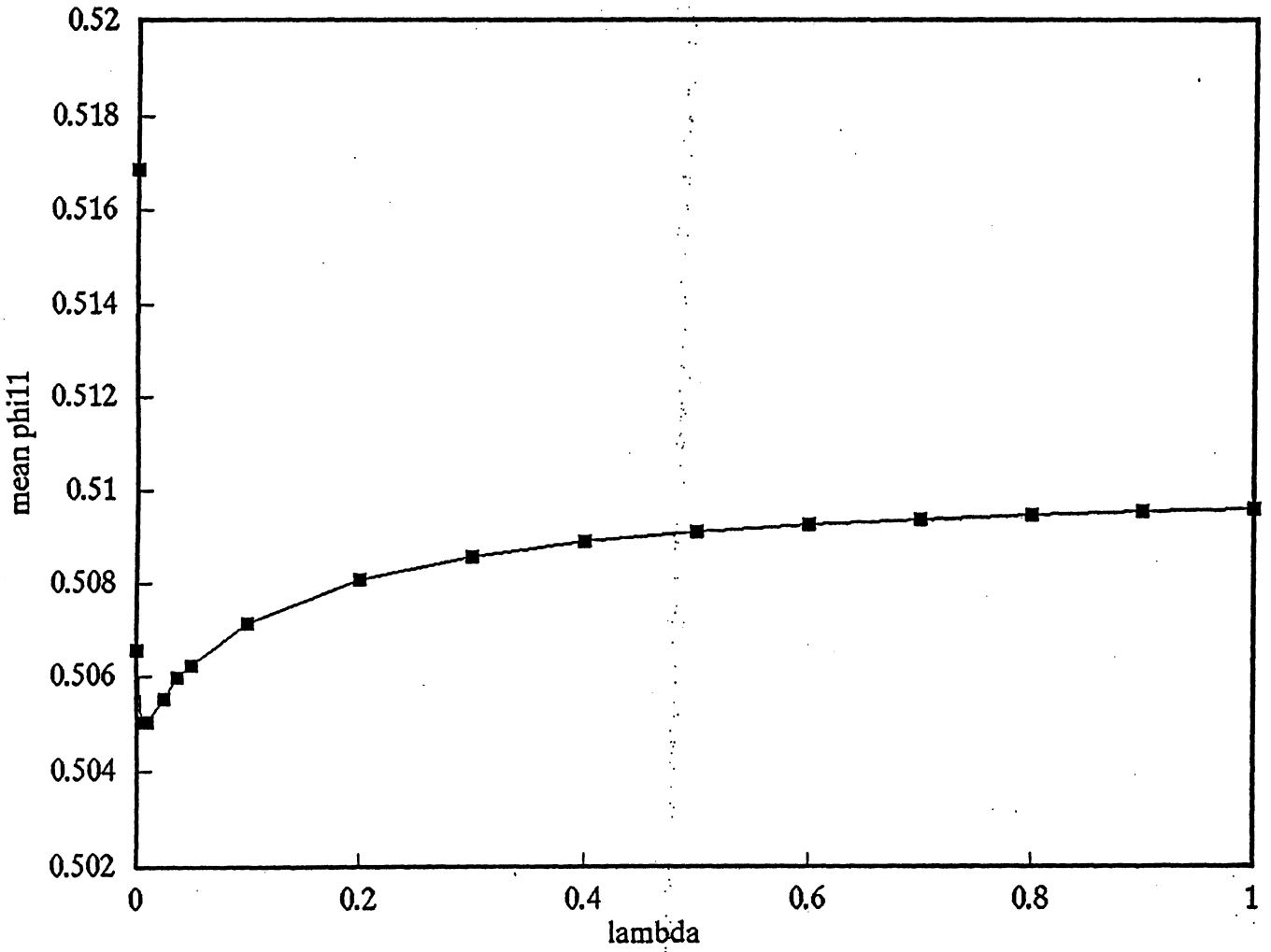

pij
0,12 0,16 0,20 0,24 0,28 0,32
j
4,8 4,4 4,0 3,6 3,2 2,8 2,4 2,0 1,6 1,2
i
1,2 1,6 2,0 2,4 2,8 3,2 3,6 4,0 4,4 4,8

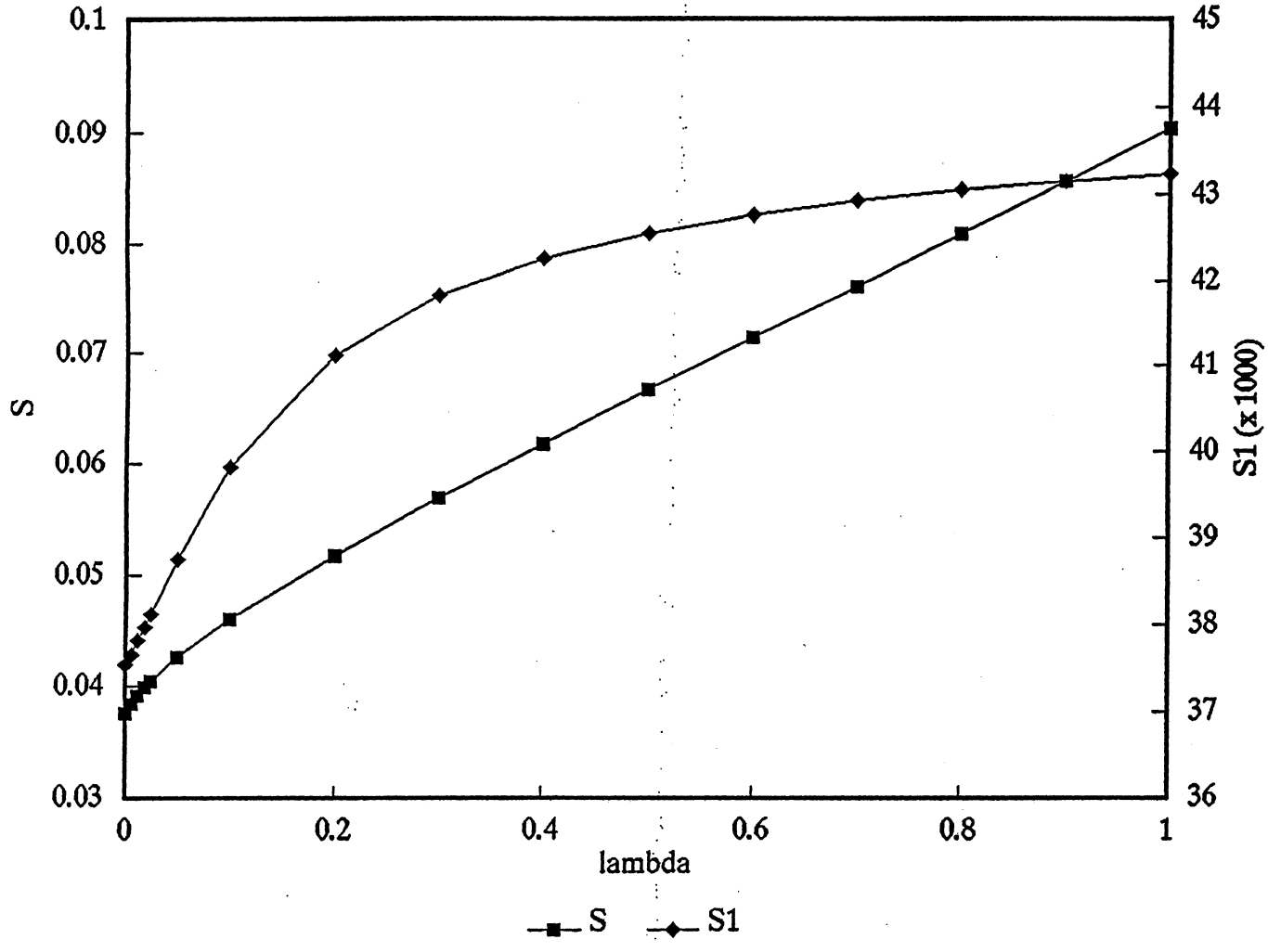

pij
0.15 0.20 0.25 0.30 0.35 0.40 0.45
j
4.8 4.4 4.0 3.6 3.2 2.8 2.4 2.0 1.6 1.2
i
1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8

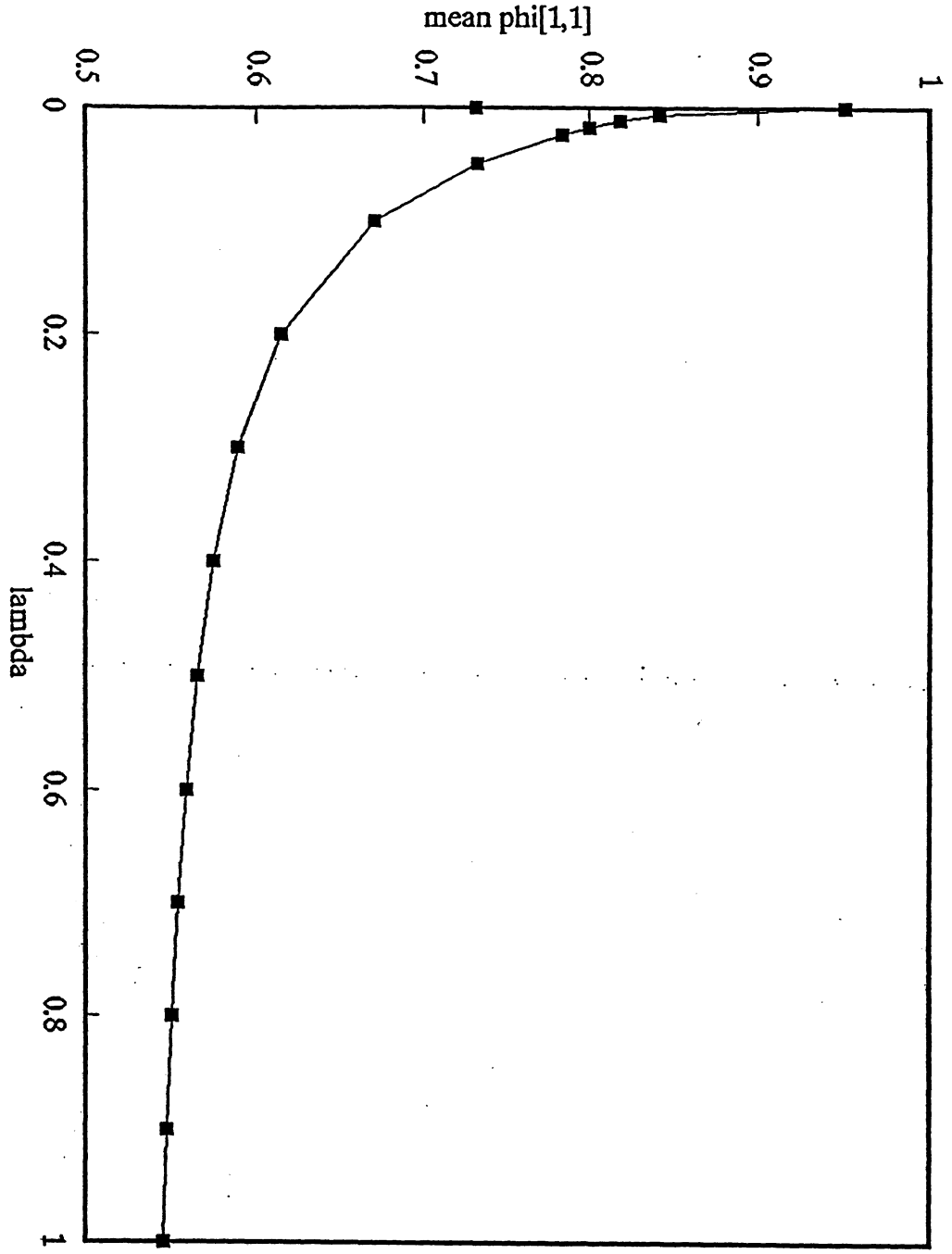

pij
0.150.200.250.300.350.400.450.50
4.8 4.4 4.0 3.6 3.2 2.8 2.4 2.0 1.6 1.2
j
1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.4 4.8
i

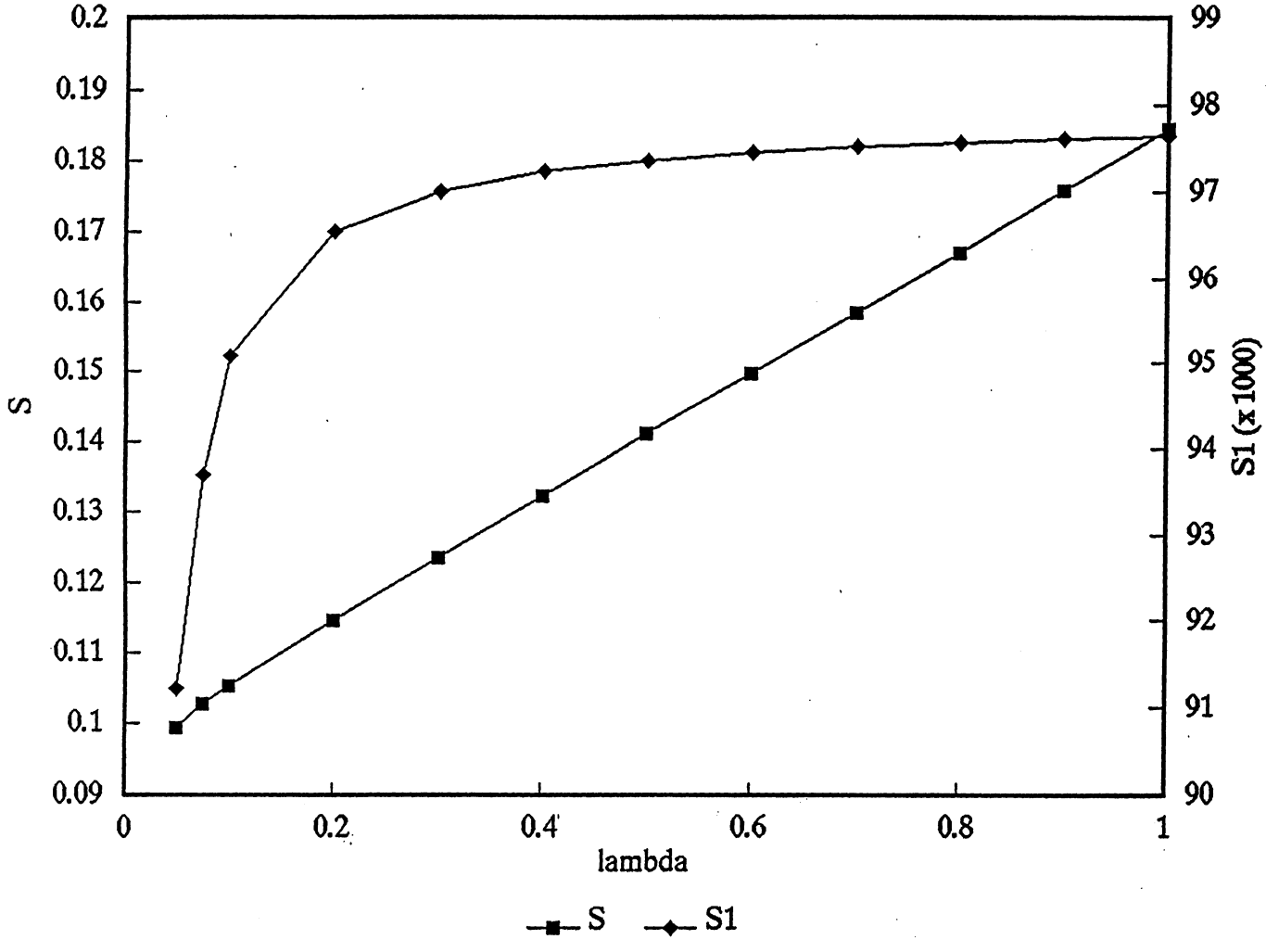

pij
0,0 0,1 0,2 0,3 0,4 0,5 0,6
j
0,4 0,8 1,2 1,6 2,0 2,4 2,8 3,2 3,6 4,0 4,4 4,8
i
1,2 1,6 2,0 2,4 2,8 3,2 3,6 4,0 4,4 4,8

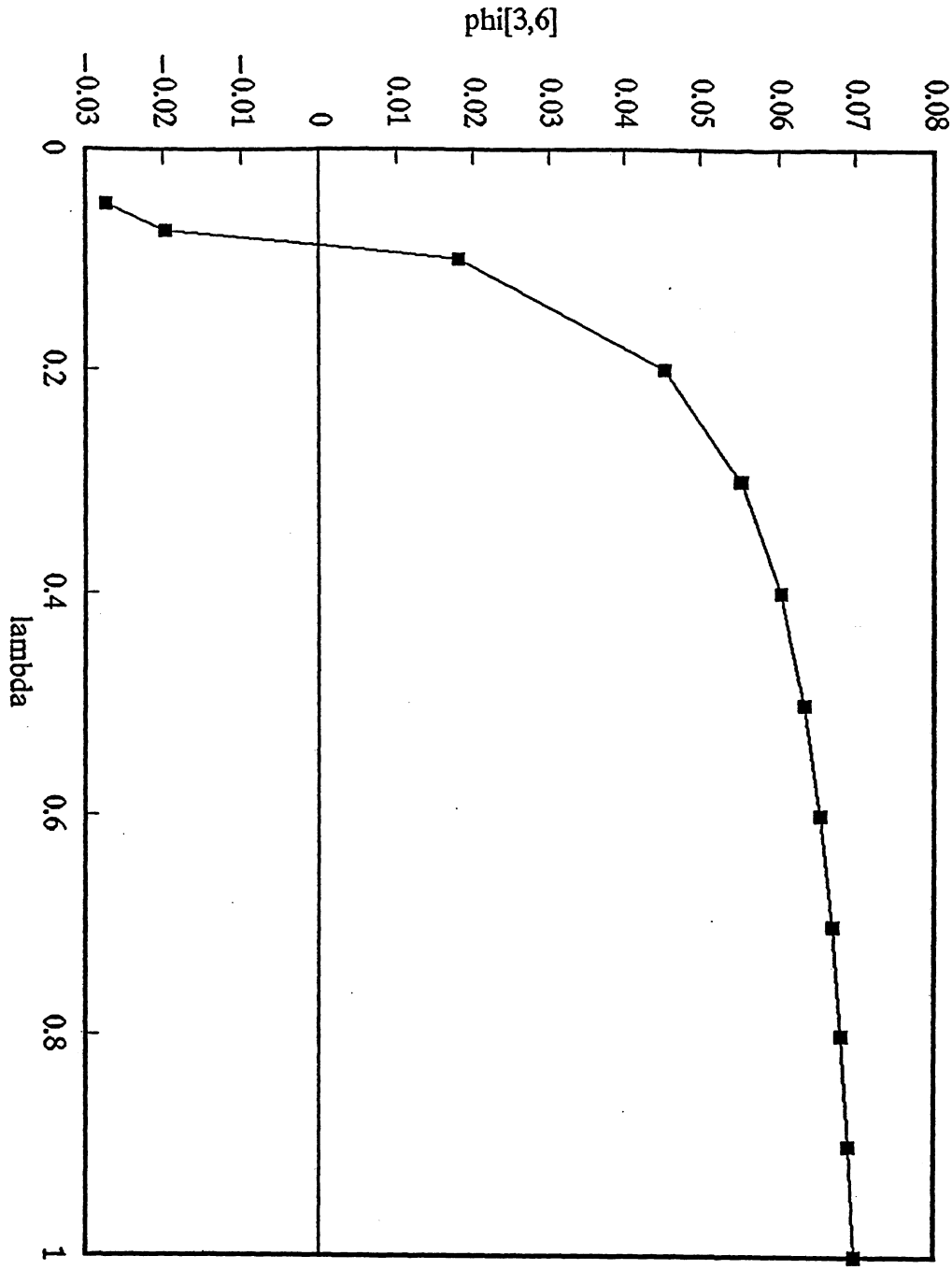

11
10.5
10
9.5
9
8.5
8
S (x 1000)
8.785
8.78
8.775
8.77
8.765
8.76
8.755
8.75
8.745
S1 (x 1000)
0
0.2
0.4
0.6
0.8
1
lambda
S
S1


0.52
0.518
0.516
0.514
0.512
0.51
0.508
0.506
0.504
0.502
mean phi11
0
0.2
0.4
0.6
0.8
1
lambda


0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
S
45
44
43
42
41
40
39
38
37
36
S1 (x 1000)
0
0.2
0.4
0.6
0.8
1
lambda
S
S1


mean phi[1,1]
0.5
0.6
0.7
0.8
0.9
1
lambda
0
0.2
0.4
0.6
0.8
1


0.2
0.19
0.18
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.1
0.09
S
99
98
97
96
95
94
93
92
91
90
S1 (x 1000)
0
0.2
0.4
0.6
0.8
1
lambda
S
S1


phi[3,6]
−0.03
−0.02
−0.01
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
lambda
0
0.2
0.4
0.6
0.8
1